DOCUMENT RESUME

ED 322 200 TM 015 402

AUTHOR Pollack, Judith M.
TITLE Some Issues in Free Response Testing.
PUB DATE 12 Jun 90
NOTE 10p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Costs; Educational Assessment; Elementary Secondary Education; Equated Scores; *Longitudinal Studies; Multiple Choice Tests; National Programs; *National Surveys; Scoring; Test Construction; *Test Format; Testing Programs; Test Reliability; Test Validity
IDENTIFIERS *Free Response Test Items

ABSTRACT

This paper summarizes an investigation of applications and issues in free response (FR) testing during 1989. It draws on ideas from the results of the National Educational Longitudinal Study 1988 (NELS:88) field test, a seminar series at the Educational Testing Service (ETS), working papers prepared for several FR testing applications, and conversations with several individuals at the ETS and elsewhere who have been implementing FR testing. Focus is on incorporating FR items in the NELS:88 test battery. Benefits of the FR format, scoring considerations, resource costs, domain coverage, reliability considerations, validity considerations, equating issues, and implications for longitudinal testing are considered. The primary advantage of the FR format is that it offers the opportunity to assess what the student can do, rather than only what the student knows. However, given the short testing time available for the NELS:88 survey, the use of FR items would necessarily entail a very sparse sampling of the content domains being tested. The resulting low reliability of the change score measurements would limit the usefulness of the data in identifying factors related to educational achievement. It is concluded that FR tests may be more accurate than multiple-choice tests, but they involve considerable cost and sacrifice reliability. The appropriate choice or mix of FR and multiple-choice testing modes depends on the purposes and uses of the test scores. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JUDITH M. POLLACK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED322200

SOME ISSUES IN FREE RESPONSE TESTING

Judith M. Pollack

TM015402

BEST COPY AVAILABLE

Some Issues in Free Response Testing

Judith M. Pollack
June 12, 1990

There is currently a great deal of interest in the use of free response items in testing because of the ways in which these items can provide information not tapped by multiple choice tests. Free response, or open ended, items can require a student to organize a logical response, provide an explanation, or carry out a process or calculation. In this way, the student can demonstrate a depth of understanding or problem solving skills.

The motivation for this paper is interest in the possibility of incorporating free response items in the NELS:88 test battery. NELS:88 is a large-scale national survey of approximately 25,000 students being carried out by NORC under contract to the National Center for Education Statistics. The survey began with eighth graders in the spring of 1988, and is following the same students as they progress through high school and afterward. Students were re-surveyed in the spring of 1990, and will be contacted again in 1992 and periodically thereafter.

At three points in time, 1988, 1990, and 1992, which would be the eighth, tenth and twelfth grade years for students who stay on track, the survey includes a battery of cognitive tests. Four subject areas are tested: Reading Comprehension, Mathematics, Science, and History/Citizenship/Geography. The eighth and tenth grade tests, which have already been administered, included only multiple choice items, although a first look at incorporating free response was included in the 1989 field test prior to assembly of the 10th grade tests. Planning is currently under way for a 1991 field test that could lay the groundwork for inclusion of free response items in the twelfth grade test battery in 1992.

This paper summarizes the author's investigation of applications and issues in free response testing during the last year. It draws on ideas from the results of the NELS:88 field test, from a seminar series at Educational Testing Service, from working papers prepared for several free response testing applications, and from conversations with a number of individuals at ETS and elsewhere who have been in the vanguard in implementing free response testing. Since this investigation was undertaken for the purpose of exploring the possibility of application of free response testing to the NELS:88 survey, the orientation of this paper reflects this perspective on the issues.

Benefits of Free Response Format

The primary advantage of the free response format is that it offers the opportunity to assess what the student <u>can do</u>, rather than only what the student knows. Principles of free response testing can be applied in areas as diverse as evaluation of musical performance, visual arts portfolios, and fluency in foreign

TM015402

languages, as well as more traditional fields such as essay writing and mathematical computations.

In a typical free response item, a student might be asked to write an essay, for example, interpreting the significance of a particular historical event. Such an exercise can provide a wide variety of information about the student's competence, including knowledge of facts about the incident and related facts leading to or from it; understanding of the context or implications of the event; skill in developing and organizing a logical argument; and ability to use language effectively and correctly to present the response. The information available in this type of essay can be used for diagnostic purposes as well as for measurement. That is, it can be used to assess the strengths and weaknesses of individual students for possible remediation. Or it can help to identify common misconceptions or deficiencies among groups of students that might indicate the need for modified teaching practices or curriculum revisions.

Unlike multiple choice, where a student needs only to choose from among alternatives presented, free response format requires a candidate to produce a response based entirely on his or her own experience and/or skill. No clue is available from the response alternatives which could prompt a correct response from a student who would not have been able to produce it independently. This virtually eliminates the possibility of guessing, which is always a factor to be dealt with in multiple choice testing. Free response format also allows for the possibility of answers--correct or incorrect--which were not anticipated by the test developer. For example, a student's particular perspective on a situation may elicit a response that is just as correct as the keyed alternative, but is not one of the options presented in a multiple choice item. Or a student who arrives at an incorrect answer--in a mathematical computation, for example--is not prompted to try again by failing to find his or her answer as an alternative.

A characteristic of assessment instruments that may not always be on the minds of test developers is that tests exist as part of a feedback loop. Tests function not only as measurement instruments, with data going from the test taker to the test user. Information goes in the other direction as well. Students--and teachers--get cues as to what is important to learn from what is measured on standardized tests. To the extent that tests require competence in problem solving, applying knowledge, demonstrating understanding, rhetorical skills and verbal expression, these will be valued and emphasized in the classroom.

Scoring Considerations

Standardized free response testing presents a variety of complications in scoring. Well written multiple-choice items have one right answer, and all others are unambiguously wrong. But a free response essay, for example, the hypothetical historical interpretation described above, is much more complex. Just as a

student can put a variety of ideas into an essay, so too is it possible to define the quality of a response in a variety of different ways. Responses to a particular question might be scored according to very different criteria, depending on the purpose of the item. Perhaps the intention is to demonstrate the student's knowledge of historical events, and writing style is not to be counted in the score. Or the student's ability to organize a convincing argument might be the primary objective of measurement, with individual facts considered of minor or no importance. The essay might be scored holistically--that is, with a range of scores assigned according to overall quality or impression--or analytically, with points awarded or deducted for the presence or absence of each of a number of characteristics. It is important to remember that free response tests do not automatically have a well-defined score scheme, and that defining objectives and developing a scoring guide accordingly are intrinsic parts of the test construction process.

It is essential for the test taker to understand the scoring criteria being used in order to target his or her response appropriately. This may be reasonably easy to do in a classroom situation, where the test developer/scorer--the teacher--and the test taker know each other. The student can draw on past experience to know what the teacher expects and values. And the teacher can evaluate the student's response in the light of other information about his or her past performance. A national standardized-testing situation does not function in this manner. The test developers, test takers, and scorers are strangers to each other, and have no information to guide their responses and evaluations other than what is explicitly provided. It is essential that scoring criteria for each exercise be clearly understood by the test takers.

Resource Costs

The advantages of the versatility and unstructured nature of free response test items come at considerable cost. The time it takes for a student to complete a free response exercise, relative to a multiple choice item, and the time, training and expertise required to score the responses, impose severe constraints on their use. A typical multiple choice item on a standardized test might be allotted a half minute to one minute for a response. But a fairly simple free response item, requiring several sentences to answer, or perhaps a mathematical computation with a diagram, takes much longer. If the student must organize and present the response in writing, not merely choose among alternatives, a minimum of possibly five to ten minutes per item is required for the simplest item types, while a more complex essay might require a half hour or more.

Responses which take longer to produce also take longer to score. Free response items must be evaluated by human beings, not by computers. Readers who are experts in the material being tested--often teachers, for tests of high school students--must be

trained in application of the scoring criteria. Prior to an operational essay reading at ETS, trainers review hundreds of student responses and select papers that exemplify the steps of the score scale. They then teach the readers to identify the points to be considered in assigning scores. For most operational tests at ETS, papers are scored by two readers independently, with a procedure specified for resolution of discrepancies. The cost of assembling hundreds of readers from across the country for a major test administration--including transportation, housing, meeting rooms, etc., as well as salary--is prohibitive for many applications. Although a trained reader can score short free response items at the rate of as much as one a minute, a computerized optical scanner can process many thousands of multiple choice items in the same amount of time. It is important to weigh the administration and scoring costs against the benefits to be gained when deciding whether free response format is appropriate for a particular application.

Domain Coverage

In a testing situation such as the NELS:88 survey, where a total of only 90 minutes is available for testing four content areas (Reading Comprehension, Mathematics, Science, and History/Citizenship/Geography), the time requirements of free response items come at considerable cost in terms of domain coverage. The relevant domain in any content area consists of all the possible knowledge or skills in that area. Any test must necessarily sample a small part of the domain of knowledge. Ideally, the sampling on a particular test needs to be representative of the domain of knowledge in order for a test score to have meaning in relation to the domain. When we say on the basis of test results that a student has mastered certain skills in mathematics, for example, we would like the test score to be a good indication of how the student might perform on the many questions that weren't asked, as well as the few that were. Obviously, the more questions included in a test, the greater the potential for adequate sampling of the relevant domain. Because of time and cost considerations described above, free response items necessarily trade off breadth of domain coverage for depth of information to be obtained from the items selected.

Reliability Considerations

A good definition of reliability for this discussion is consistency of measurement. That is, a test-taker should ideally receive the same score on a different set of items sampled from the same domain, or on the same set when scored by a different reader. The latter requirement seems not to be a serious drawback of a well designed free response scoring scheme. When scoring guides are carefully written and readers trained to apply them accurately, reader agreement tends to be extremely high, generally exceeding 90%. Although scoring reliability imposes an upper limit on the reliability of a test, it does not necessarily compromise the overall reliability to an unacceptable extent.

The potential for score discrepancy due to the choice of items on the test is a much more serious problem. The content, context, or cognitive demands of any single item interact with a particular test taker's personal characteristics and experience in ways that are not related to the construct being measured. Presenting a large number of different items in a test is a good way to reduce measurement errors due to irrelevant differences between individuals. For example, in a test of reading comprehension, reading passages with a science or technology context tend to favor individuals with experience in those areas. Even if all the information needed to answer questions is contained in the passage, and no prior knowledge of the material is required, students who have taken more science courses, and males, tend to have an advantage. So a reading comprehension test ought to balance passages with a science context along with those based on the humanities, contain materials that refer to female characters as well as males, etc. Use of free response format severely constrains the number of test items that can be administered in a given period of time. In general, a free response test can be expected to have substantially lower reliability than a multiple choice test taking the same amount of time simply because many fewer items can be administered.

Validity

The validity of a test can be defined as the extent to which a test is doing the job it is supposed to do. Two independent, but related, aspects of this idea are construct validity--whether the test is in fact measuring the characteristics it is designed to measure--and criterion, or predictive validity--whether the test score correlates with other variables in a useful manner. Construct validity requires precise definition of the domain of knowledge being measured, and judgments about appropriate sampling from the domain for the purposes of the test. Evidence from a 1989 field test of two free response items in the NELS:88 survey, one in mathematics and one in science, suggests that this item type can have a complex factor structure. In addition to the mathematical comprehension or science knowledge required, facility in verbal expression was a necessary condition for answering the item successfully. Each of the items was administered to students along with multiple choice items in two content areas, its own subject area (mathematics or science) along with one other (history or reading comprehension). In general, the free response scores were more highly correlated with the subset of multiple choice items classified as "problem solving" than with those categorized as "declarative knowledge" or "comprehension." But they also had strong relationships with multiple choice scores obtained on the second content area in the same test booklet. The science free response item had a correlation of .33 with the total science multiple choice score, and .32 with the reading score; mathematics correlated .56 with the other mathematics items, and .46 with history. This clearly indicates that factors other than math or science knowledge are contributing to the students' responses.

A third aspect of validity is believability, or face validity. While this may not be as easy to study or quantify as the others, it is at least as important. Test takers need to see test items as appropriate and relevant in order to take them seriously. And the users of test scores need to believe that the items are suitable measures of the skills being tested in order to have confidence that the results are meaningful.

Equating

Tests are typically used to measure change in individuals over time, or to make comparisons between different individuals or groups at the same time or at different times. If different test instruments are used for measurements that need to be compared, a mechanism for determining comparability of scores is required. There are a number of different approaches used for equating tests that are intended to measure the same thing but may vary in difficulty. Random samples of people from the same population may each be given a different test form, with the random assignment of forms used to justify the assumption of equal ability in the different groups. Or a group of people may each be given both of two forms to be equated, ideally with half of them receiving the forms in reversed order. Or two groups may receive two different forms, as well as a common test of "anchor" items administered along with both. Each of these methods relies on the assumption that the tests to be equated are measuring the same thing.

Equating of free response tests presents some unique problems. Because of the complex factor structure of free response items, it is hard to define tasks that are of equal difficulty. A particular aspect of a task may be more challenging for one individual and less so for another, while some other requirement of the test item may reverse the rank-ordering of the same test takers. Comparable topics are as difficult to define as comparable difficulty, again because of each test taker's unique background and experience.

Assuming difficulty could be defined in the same way for two free response tasks, the typical short score scale presents a second problem. A free-response item scored holistically might be scored on a fairly short scale, say 1 to 5. If another item to be equated to it turned out to be slightly more difficult for the same population, it could be determined that a score of 3 on the first item is equivalent to a score somewhere between 2 and 3 on the second. This coarse equating is not likely to be satisfactory, but assigning a more precise number, perhaps 2.8, implies much more precision in the measurement process than actually exists.

Test security, in many situations, can present a third problem. Equating frequently entails re-use of anchor items to tie one test to another. While a test taker is unlikely to remember very many of say, 100 multiple choice items after completing a one-hour test, he or she is extremely likely to remember the topics of one or two essay questions administered in

the same time. In situations where prior knowledge of test questions is of considerable value to individuals, such as admissions or qualification tests, free response items cannot be re-used. Thus a valuable equating tool is not available.

Implications for Longitudinal Testing

Experience with previous NCES longitudinal studies indicates that the NELS:88 survey data will be used extensively to study many aspects of students' educational experiences and outcomes. Policy studies will focus on determining what types of influences lead to achievement gains for students during the high school years. Small subgroups of students are likely to be studied as well as the entire sample. Measurement of change over time requires scores that are highly reliable at the individual level.

Another major consideration in a longitudinal study that spans a substantial period of time is equating of scores. The NELS:88 tests are administered in eighth, tenth, and twelfth grades. In order to avoid floor effects at the eighth grade, and ceiling effects later on, the tests are designed to be grade-level adaptive. That is, the difficulty range of the test items increases from year to year to keep pace with students' improved skills, while a core number of items are retained in each test for equating purposes. The IRT (Item Response Theory) equating procedure being utilized to establish a common scale for measuring change across years is applicable only to single-factor subtests, with items scored right/wrong.

Given the short testing time available for the NELS:88 survey, use of free response items would necessarily entail a very sparse sampling of the content domains being tested. The resulting low reliability of the change score measurements would limit the usefulness of the data in indentifying factors related to educational achievement.

Ideally, for any test, we would like to be able to measure the right thing, and measure it well. Most tests may be a compromise between these two ideals. Multiple-choice tests can be an efficient and reliable measurement tool, but may not address the full range of skills and knowledge that are relevant to the test's objectives. Free response tests may be able to come closer to measuring the "right thing", but necessarily do so at considerable cost, and sacrifice of reliability. The appropriate choice or mix of these two testing modes depends on the purposes and uses of the test scores.

References

Educational Testing Service Seminar Series:

Issues Involved in Developing and Scoring Free Response Tests, M. Fowles and C. Solomon, December, 1989.

Measurement Issues in Free Response Testing, S. Livingston, January, 1990.

ETS Core Scoring, G. Conlan, April, 1990.

A Question of Thinking: A First Look at Students' Performance on Open-ended Questions in Mathematics, California State Department of Education, Sacramento, 1989.

Personal Communication: G. Conlan, S. Livingston, G. Conlan, T. Pandey, M. Fowles, K. Yamamoto.